VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF INFORMATION TECHNOLOGY

FINAL PROJECT REPORT

CS313.P11.KHTN - DATA MINING AND APPLICATION

Adversarial Detection with Medical Model via Hypothesis Test

Instructor: Dr. Vo Nguyen Le Duy

Students of Group 5:

Full name	Student ID
Tran Nhat Khoa	22520691
Huynh Tong Dang Khoa	22520670
Le Tran Quoc Khanh	22520638

Ho Chi Minh City, December 2024

Acknowledgments

We would like to express our deep gratitude to Dr. Vo Nguyen Le Duy for his dedicated teaching and valuable guidance in the Data Mining and Application course. Your support and feedback have been instrumental in shaping our project. Despite our dedicated efforts to complete this project, we recognize the possibility of overlooked errors. We genuinely appreciate your constructive feedback to help us refine and improve further.

Task Allocation Table

Full name	**Student ID**	**Workload**
Tran Nhat Khoa	22520691	40%
Huynh Tong Dang Khoa	22520670	30%
Le Tran Quoc Khanh	22520638	30%

1 Introduction

1.1 Motivation

Convolutional Neural Networks (CNNs) are widely used in medical imaging but are vulnerable to adversarial attacks, which pose serious risks to patient safety. Existing defenses, such as adversarial training and denoising, often alter model weights or distort critical details, making them unsuitable for clinical settings. Thus, it is vital to prevent adversarial attacks rather than defend against them.

1.2 Problem Statement

- **Input:**
  - f : A classification model that also serves as a feature extractor, mapping input data to a d-dimensional feature space.
  - X : The original dataset used to train f.
  - Y : A new image dataset, which may contain adversarial examples.
- **Output:** A decision indicating whether Y contains adversarial examples, based on the feature distribution of X and Y.

2 Method

To detect adversarial examples, we measure the distance between the features extracted from the training dataset X and the test dataset Y. A function f_X, trained on X, maps the data to a feature space where distributional differences caused by adversarial attacks can be highlighted. This method focuses on high-level, semantic representations extracted from f_X, which are robust to minor adversarial perturbations.

The detection process involves transforming the datasets into their feature representations, computing the distance between these representations, and applying a threshold-based decision rule. The steps are formalized in Algorithm 1.

Algorithm 1 Adversarial Detection using Feature Distribution Distance

```
Input: Datasets $X = \{x_1, x_2, \ldots, x_n\}$, $Y = \{y_1, y_2, \ldots, y_m\}$, function $f_X$ trained in $X$,
       threshold $t$
Output: Two distributions are the same or not
Step 1: Apply function $f$ to datasets  $f_X(X) = \{f_X(x_1), f_X(x_2), \ldots, f_X(x_n)\}$  $f_X(Y) =$
  $\{f_X(y_1), f_X(y_2), \ldots, f_X(y_m)\}$
Step 2: Compute Distribution distance for the transformed data  Compute distribution
  distance between $f_X(X)$ and $f_X(Y)$ using the formula:
      $\text{Dis} = \text{calDistributionDistance}(f_X(X), f_X(Y))$
  Step 3: Decision rule using threshold
  if $Dis > t$ then
      Distributions are different
else
      Distributions are the same
```

To measure the distance between the feature distributions $f_X(X)$ and $f_X(Y)$, we use the **Maximum Mean Discrepancy (MMD)**. The MMD is a non-parametric metric that does not require prior knowledge of the specific form of the underlying data distributions, making it highly flexible and robust for adversarial detection tasks. Formally, given two sets of feature representations $f_X(X)$ and $f_X(Y)$, the squared MMD is defined as:

$$\text{MMD}^2(X, Y) = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} k(x_i, x_j) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j),$$

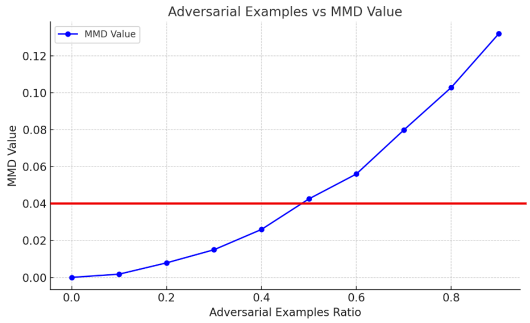
where $k(x, y)$ is the kernel function chosen as the Gaussian radial basis function (RBF):

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right).$$

Here, m and n are the sizes of the two sets X and Y, and σ is the bandwidth parameter of the kernel.

One of the key advantages of using the MMD is its non-parametric nature: the MMD does not require prior knowledge about the form or structure of the underlying distributions. This makes it suitable for comparing complex or unknown distributions, such as those resulting from adversarial manipulations.

As shown in Figure 1, the MMD distance between the adversarial attack dataset and the clean dataset is significantly higher than other non-toxic transformed datasets.



Transform method	MMD
Flips (100%)	0.00049
Gaussian Blurring (100%)	0.00883
Flips (50%) + Subsampling (50%)	0.02011
Subsampling (100%)	0.03122
FGSM (50%)	0.04253
PGD (50%)	0.02889
FGSM (100%)	**0.16246**
PGD (100%)	**0.12085**

Figure 1: Comparison of MMD values: (a) Graph showing MMD values and Adversarial Examples Ratio, (b) Table summarizing MMD values for various transformation methods.

We can also notice that the MMD value between the adversarial dataset and the transformed non-toxic (flips, supersampling) dataset may be approximately equal.

Thresholding provides a rigid decision rule without indicating the level of certainty behind that decision. However, using a fixed threshold to classify adversarial examples may not capture the nuances in the distribution shift.

3 Hypothesis Testing with MMD and Bootstrapping

To solve the problem above, we proposed a method for hypothesis testing using Maximum Mean Discrepancy (MMD) combined with a bootstrap approach. The method aims to identify whether two datasets, such as clean and adversarial samples, come from the same distribution. By leveraging MMD, a non-parametric measure, we quantify the distributional difference between the two datasets. To establish the statistical significance of the result, we perform bootstrapping to estimate the p-value. This approach ensures reliable detection of adversarial examples even with small sample sizes. The following algorithm summarizes the proposed hypothesis testing framework:

Algorithm 2 Hypothesis Testing with MMD and Bootstrapping

```
Input: Datasets $X = \{x_1, x_2, \dots, x_n\}$, $Y = \{y_1, y_2, \dots, y_m\}$
Output: p-value for hypothesis test
Step 1: Define hypotheses
  Null Hypothesis: $H_0 : P_X = P_Y$, i.e., $X$ and $Y$ come from the same distribution
  Alternative Hypothesis: $H_1 : P_X \neq P_Y$, i.e., $X$ and $Y$ come from different distributions
Step 2: Compute MMD for the original data
  $$\text{MMD}_{ob} = \text{MMD}(f_X(X), f_X(Y))$$
Step 3: Generate bootstrap samples for each iteration $i = 1$ to $N_{bootstrap}$ do
    Sample bootstrap datasets $X^* = \{x_1^*, \dots, x_n^*\}$ and $Y^* = \{y_1^*, \dots, y_m^*\}$ from $X$ and $Y$
Step 4: Compute MMD for bootstrap samples for each bootstrap sample pair $(X^*, Y^*)$ do
    Compute MMD for each pair $(X^*, Y^*)$:
    $$\text{MMD}_{bo} = \text{MMD}(f_X(X^*), f_X(Y^*))$$
Step 5: Create MMD distribution from bootstrap samples Repeat Steps 3 and 4 for $N_{\text{bootstrap}}$ iterations to obtain a distribution of bootstrap MMD values:
  $$\{\text{MMD}_{bo,1}, \text{MMD}_{bo,2}, \dots, \text{MMD}_{bo,N_{\text{bootstrap}}}\}$$
Step 6: Calculate p-value
  $$p = \frac{|\{\text{MMD}_{bo,i} \geq \text{MMD}_{ob}\}|}{N_{\text{bootstrap}}}$$
Step 7: Decision Rule if $p < \alpha$ then
    Reject $H_0$: Conclude that $X$ and $Y$ come from different distributions.
else
    Fail to reject $H_0$: Conclude that there is insufficient evidence to claim a difference between $X$ and $Y$.
```

4 Hypothesis Test Results

The results for various transformation methods are summarized in Table 1. The MMD values and p-values are used to detect whether the transformed datasets are adversarial.

Table 1: Hypothesis Test Results

Method	MMD	p-value	Detected/Target
Flips (100%)	0.00049	1.000	Clean / Clean
Gaussian Blurring (100%)	0.00883	0.998	Clean / Clean
Flips (50%) + Subsampling	0.02011	1.000	Clean / Clean
Subsampling (100%)	0.03122	0.822	Clean / Clean
FGSM (50%)	0.04253	0.000	Adversarial / Adversarial
PGD (50%)	0.02889	0.000	Adversarial / Adversarial
FGSM (100%)	0.16246	0.000	Adversarial / Adversarial
PGD (100%)	0.12085	0.000	Adversarial / Adversarial

As from the result, the proposed Hypothesis Testing with MMD and Bootstrapping performed effectively in distinguishing clean data from adversarial samples.

- For clean transformations (e.g., Flips, Gaussian Blurring), the p-values are close to 1.0, indicating no significant difference, as expected.
- For adversarial methods (e.g., FGSM, PGD), higher MMD values and p-values close to 0 demonstrate the method's ability to detect significant distributional shifts.

5 Experimental Setup

The experiments are conducted on the **Brain Tumor MRI** dataset, with the test set containing 1300 images. Clean images and adversarial examples are analyzed to measure the ability of the hypothesis tests to distinguish between the two.

5.1 Adversarial Attacks

- **FGSM**: Fast Gradient Sign Method (FGSM), formulated as:

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)). \tag{1}$$

- **PGD**: Projected Gradient Descent (PGD), iteratively applied as:

$$x_{\text{adv}}^{(t+1)} = \text{Proj}_{x+B_\epsilon}\left(x_{\text{adv}}^{(t)} + \alpha \cdot \text{sign}(\nabla_x J(\theta, x_{\text{adv}}^{(t)}, y))\right). \tag{2}$$

5.2 Classification Model - Feature Extractor

We use a ResNet model trained on a clean dataset of 8000 images, achieving high performance on the clean data.

Table 2: Performance of Classification Model under Different Attacks.

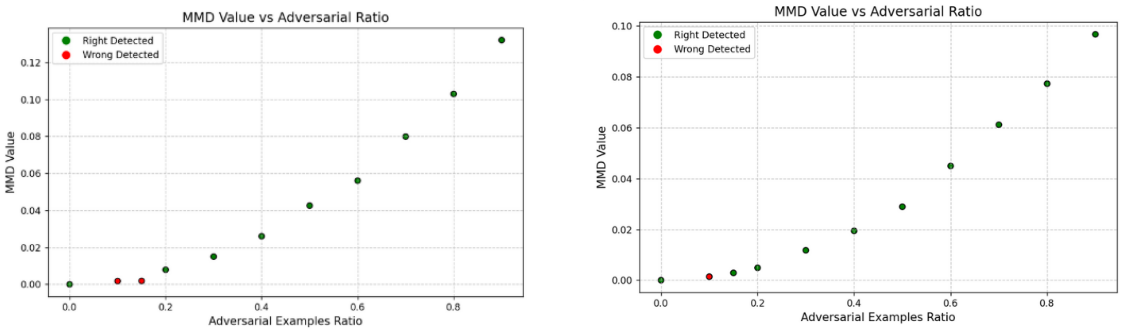
Datasets	No Attack	DDN	PGD	FGSM
Brain Tumor MRI	**0.9329**	0.5812	0.7880	0.7101

5.3 Hypothesis Testing

We apply a hypothesis test using algorithm 2 to detect adversarial examples. The p-values from the test indicate whether adversarial samples are reliably detected.

6 Results

Our proposed method demonstrates strong detection performance, with results improving as the proportion of adversarial examples increases. Reliable detection is consistently achieved when at least 20% of the dataset comprises adversarial examples, as shown in Figures 2a and 2b.



(a) FGSM

(b) PGD

Figure 2: Comparison of MMD values for FGSM and PGD adversarial ratios.

Smaller test sizes negatively impact the statistical power, reducing the reliability of p-values. Figure 3 highlights this effect, emphasizing the need for sufficient sample sizes to ensure robust performance.

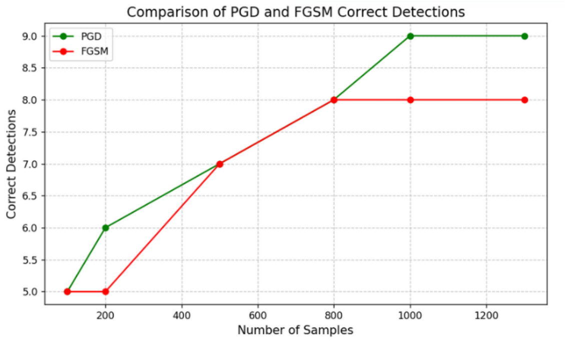


Figure 3: Effect of Test Size on Detection Accuracy.

The results suggest that as the proportion of adversarial examples increases, the MMD value rises, and the p-value approaches 0.000, leading to reliable detection. For both FGSM and PGD attacks, a dataset with at least 20% adversarial examples is sufficient for the hypothesis test to reliably distinguish them.

7 Conclusion

In this project, we developed an adversarial detection method for medical models using hypothesis testing with Maximum Mean Discrepancy (MMD). By leveraging feature distribution analysis and bootstrap-based hypothesis testing, we achieved robust detection of adversarial examples with high statistical reliability. Experimental results demonstrated the effectiveness of the proposed approach, particularly in distinguishing adversarial datasets from clean datasets. This method offers a non-parametric, scalable solution for enhancing the security of medical imaging systems against adversarial attacks.

The implementation and code for reproducing our experiments can be found in the following GitHub repository: GitHub.